迁移学习研究进展

庄福振　　何 清

摘要 近年来，迁移学习已经引起了广泛的关注。迁移学习是运用已存有的知识对不同但相关领域问题进行求解的新的一种机器学习方法。传统机器学习基于两个基本假设：(1) 用于学习的训练样本与新的测试样本满足独立同分布的条件；(2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型。迁移学习降低了要求，目的是迁移已有的知识来解决目标领域中仅有少量或没有有标签样本数据时的学习问题。本文对迁移学习算法以及相关理论研究进展进行了综述，并介绍了我们在该领域所做的研究工作，特别是利用生成模型在概念层面建立迁移学习模型。最后指出了迁移学习下一步可能的研究方向。

关键词 迁移学习，独立同分布，生成模型

1　引言

随着社会发展的信息化和网络化，人们在日常生活和工作中无时无刻不在获取信息，分析信息，并以此作为决策的依据。在一定程度上，信息的拥有量已经成为决定和制约人类社会发展的重要因素。想要高效准确地寻找到所需的信息，信息分类是必不可少的第一步。通过分类，信息可以得到有效的组织管理，有利于快速准确的定位信息。分类学习问题，是机器学习中一种重要的学习方法，目前已经得到广泛的研究与发展。

在传统分类学习中，为了保证训练得到的分类模型具有准确性和高可靠性，都有两个基本的假设：(1) 用于学习的训练样本与新的测试样本满足独立同分布的条件；(2) 必须有足够可利用的训练样本才能学习得到一个好的分类模型。但是，在实际应用中我们发现这两个条件往往无法满足。首先，随着时间的推移，原先可利用的有标签的样本数据可能变得不可用，与新来的测试样本的分布产生语义、分布上的缺口。比如，股票数据就是很有时效性的数据，利用上月份的训练样本学习得到的模型并不能很好地预测本月份的新样本。另外，有标签的样本数据往往很缺乏，而且很难获得。在 Web 数据挖掘领域，新数据不断涌现，已有的训练样本已经不足以训练得到一个可靠的分类模型，而标注大量的样本又非常费时费力，而且由于人的主观因素容易出错。这就引出了机器学习中另外一个重要问题，如何利用少量的有标签训练样本或者源领域数据，建立一个可靠的模型，对目标领域数据进行预测（源领域数据和目标领域数据可以不具有相同的数据分布）。何清等人[1]指出数据分类首先要解决训练集样本抽样问题，如何抽到具有代表性的样本集作为训练集是一个值得研究的重要问题。该文提出极小样本集抽样方法，用于基于超曲面分类算法。该方法可感知非结构化数据的分布，并以极小样本集作为代表子集。该文还指出了极小样本集有多少种表达方式。给出了样本缺失情况下准确率的精确估计。这篇文章表明在实际中保证训练得到的分类模型具有高准确性和可靠性的两个基本的假设并不是每个算法都能做到的，因此研究迁移学习变得非常重要。

近年来，迁移学习已经引起了广泛的关注和研究[2-18]。根据维基百科的定义[1]，迁移学习是运用已存有的知识对不同但相关领域问题进行求解的新的一种机器学习方法。它放宽了传统机器学习中的两个基本假设，目的是迁移已有的知识来解决目标领域中仅有少量或甚至没

[1] http://en.wikipedia.org/wiki/Transfer_learning

有有标签样本数据时的学习问题。迁移学习广泛存在于人类的活动中。两个不同的领域共享的因素越多，迁移学习就越容易，否则就越困难，甚至出现“负迁移”[19-20]，产生副作用。比如：一个人要是学会了骑自行车，那他就很容易学会开摩托车；一个人要是熟悉五子棋，也可以轻松地将知识迁移到学习围棋中。但是有时候看起来很相似的事情，却有可能产生“负迁移”。比如，学会骑自行车的人来学习骑三轮车反而不适应，因为它们的重心位置不同[21-22]。近几年来，已经有相当多的研究者投入到迁移学习领域中，每年在机器学习和数据挖掘的顶级会议（如 ICML、SIGKDD、NIPS、ICDM、CIKM 等）中都有关于迁移学习的文章发表。

2 迁移学习算法研究进展

针对源领域和目标领域样本是否标注以及任务是否相同，可以把以往迁移学习工作划分为归纳迁移学习、直推式迁移学习以及无监督迁移学习等[13]。而按照迁移学习方法采用的技术划分，又可以把迁移学习方法方面的工作大体上分为：i) 基于特征选择的迁移学习算法研究；ii) 基于特征映射的迁移学习算法研究；iii) 基于权重的迁移学习算法研究。本文主要从这两条线对迁移学习的工作进行总结和综述。

2.1 按源领域和目标领域样本是否标注以及任务是否相同划分

潘嘉林(音译，S. J. Pan)和杨强(音译，Qiang Yang) [13]针对源领域和目标领域样本是否标注以及任务是否相同或者是否单一对迁移学习进行了划分，如图 1 和表 1 所示。

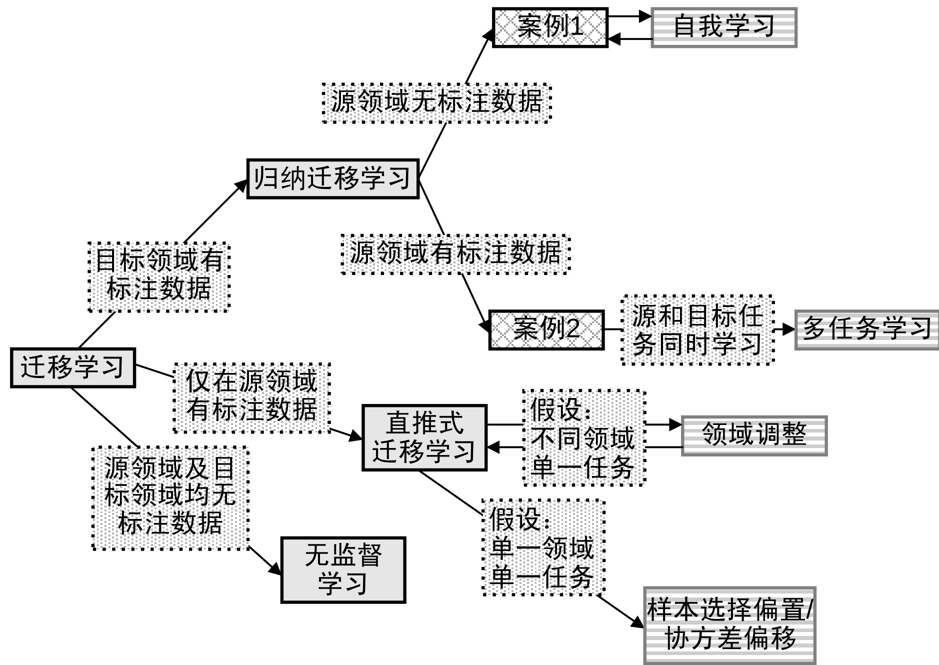


图1. 迁移学习按照源领域和目标领域样本是否有标注进行划分[13]

从图 1 可以看到，根据源领域和目标领域中是否有标签样本，迁移学习可划分为三类：目标领域中有少量标注样本的归纳迁移学习（Inductive Transfer Learning）、只有源领域中有标签样本的直推式迁移学习（Transductive Transfer Learning）以及源领域和目标领域都没有标签样本的无监督迁移学习。另外根据源领域中是否有标签样本，还可以把归纳迁移学习划分成多任务学习、自学习。表 1 给出了传统机器学习与各种迁移学习之间的关系，以及各种情形下，源领域与目标领域是否相同，源领域与目标领域的任务是否相同。迁移学习是和传

统学习相对应的一大类学习方式，传统学习处理源领域和目标领域相同且源领域和目标领域的任务也相同的学习；迁移学习处理除此情形之外的学习，包括：源领域和目标领域的任务相关但不同的归纳迁移学习[6][12][23-28]；源领域和目标领域相关但不相同而源领域和目标领域的任务相同的直推式迁移学习（Transductive Transfer Learning）[3][7][29-33]。无监督迁移学习与归纳迁移学习类似，不过主要处理源领域和目标领域中都没有标签数据的问题[34-35]。还根据训练样本和测试样本是否来自于同一个领域，把直推式迁移学习划分为样本选择偏差、协方差偏移和领域自适应学习这些相关的子领域。

表 1. 传统机器学习与各种迁移学习情形之间的关系[13]

学习设置		源领域和目标领域	源任务和目标任务
传统机器学习		相同	相同
迁移学习	归纳迁移学习	相同或者不同但相关	不同但相关
	无监督迁移学习	相同或者不同但相关	不同但相关
	直推式迁移学习	不同但相关	相同

2.2 按采用的技术划分

下面首先介绍与迁移学习极其相关的半监督学习以及多任务学习方法，然后再对采用各种技术的工作进行介绍。

2.2.1 半监督学习方法

在传统的监督学习中，学习算法通过对大量有标签的训练样本进行学习，从而建立模型用于预测标记新来的没有标签的测试样本。但是随着信息技术、互联网以及存储技术的快速发展，数据量随时间呈指数级增长。人们能够比较容易地收集大量的没有标签的数据，但要获取大量有标签的数据则较为困难，因为这可能需要耗费大量的人力物力。例如，在生物学中进行数据分类，得到一个训练样本的标签往往需要大量的，长时间的，昂贵的实验；在进行 Web 网页推荐时，用户也不愿意花费大量的时间来标记哪些网页是他感兴趣的，因此有标签的网页很少。实际上，在真实世界中通常存在大量的无标签的数据，而有标签的数据则较少。这就需要一种机器学习技术，能够利用大量的无标签样本数据以及少量有标签的训练样本进行学习，提高分类任务的准确率。

按照周志华（音译，Zhi-Hua Zhou）等人[36]在文献中的阐述，目前能够利用少量有标签数据和大量没有标签样本数据的技术有三类：半监督学习(Semi-supervised Learning)、直推式学习（Transductive Learning）和主动学习（Active Learning）。这些学习方法都通过大量的无标签样本来辅助少量有标签样本的学习，学习过程中不需要人工干预。但它们的思路又有些不同。直推式学习假设无标签的数据就是最终要用来测试的数据，学习的目的就是在这些数据上取得最佳泛化能力。与之不同，半监督学习基于自身对无标签数据加以利用，在学习时并不知道最终的测试用例是什么。因此，半监督学习考虑的是一个“开放的世界”，即在学习中不知道测试样本是什么，而直推式学习考虑的则是一个“封闭世界”，要测试的样本数据已经参与到学习过程中。如果抛开是否对未知样本进行预测，其实直推式学习可以归结为半监督学习的一种特例。主动学习与半监督学习、直推式学习最大的区别在于它的学习过程需要人工的干预，就是在学习过程通过反馈尽可能地找到那些包含信息量大的样本来辅助少量有标签样本的学习。在传统机器学习中，这三种方法已经得到了广泛应用[37-42]。多视角学习（Multi-view Learning）也是半监督学习一个很重要的学习任务。雅罗沃斯基（Yarowsky）[43]和布朗姆（Blum）等人[44]认为数据的多视角表示方式可以提高半监督分类学习算法的性能。

更进一步，文献[45-47]用概率近似正确（PAC，Probability Approximately Correct，）方法分析了联合训练(Co-training)在无标签数据上错误率的上界。

近年来也有很多研究者把这些技术应用到迁移学习领域。文献[22]对主动迁移学习模型进行了研究。施潇潇（音译，Xiaoxiao Shi）等人[48]提出了一种跨领域的主动迁移学习方法，通过似然偏置的大小来选择领域外(out-of domain)有标签的样本。那些能够正确预测领域内(in-domain)数据且高似然偏置的有标签样本被利用，而那些低偏置的样本则通过主动学习进行选择。廖学军（音译，Xuejun Liao）等人[6]提出了一种方法，估计源领域中的每个样本与目标领域中少量标签数据之间的不匹配程度，并把该信息应用到逻辑回归中。庄福振等人[17]综合半监督学习的三种正则化技术——流形正则化[49]、熵正则化[50]以及期望正则化[51]，提出基于混合正则化的迁移学习方法。该方法首先从源领域训练得到一个分类器，然后通过混合正则化在目标领域数据上进行优化。

自学习(Self-taught Learning)[34][52]也是一种利用大量无标签数据来提高给定分类聚类任务性能的方法。自学习被应用于迁移学习中，因为它不要求无标签数据的分布与目标领域中的数据分布相同。瑞纳（Raina）等人 [52]提出了一种自学习的方法，它利用稀疏编码技术对无标签的样本数据构造高层特征，然后少量有标签的数据以及目标领域无标签的样本数据都由这些简洁的高层特征表示。实验表明这种方法可以极大地提高分类任务的准确率。

2.2.2 多任务学习方法

多任务学习是同时对几个相关的问题进行学习的机器学习方法。这些任务共享相同的表示。这种学习方式同样可以得到更好的模型，因为在学习中允许各个任务使用它们之间共性的东西。因此多任务学习[53-54]也可以看成是迁移学习早期的研究。卡鲁阿纳（Caruana）指出多个任务在使用共同的表示时，可以并行地执行，而且这些任务在学习中相互获利，比单个任务的学习更好。多个任务学习可以应用于许多不同的领域和不同的算法，因此在现实世界中也是非常有用的。

巴克（Bakker）等人[55]运用贝叶斯方法去估计多个问题所共有的特征参数，从而解决多任务学习的问题。白静（音译，Jing Bai）等人[56]研究学习了多个任务中的非参数共同结构，然后提出了一种算法迭代，发现对所有任务都有效的超级特征，最终生成每个任务的函数估计是这些超级特征的线性组合。文献[57]利用特征和核函数的选择结合支持向量机来解决多任务学习的问题。阿伊里乌（Argyriou）等人[58]提出了一种针对多任务的空间降维技术，试图寻找一个可以表示所有任务的低维特征空间。类似相关的工作还有引文[59-60]。但多任务学习与迁移学习不同的是，它强调算法在所有任务上都要表现得很好，而迁移学习只强调目标领域上的性能。

2.2.3 基于特征选择方法

基于特征选择的迁移学习方法是识别出源领域与目标领域中共有的特征表示，然后利用这些特征进行知识迁移[4][61-62]。蒋静（音译，Jing Jiang）等人[61]认为与样本类别高度相关的那些特征应该在训练得到的模型中被赋予更高的权重，因此他们在领域适应问题中提出了一种两阶段的特征选择框架。第一阶段首先选出所有领域（包括源领域和目标领域）共有的特征来训练一个通用的分类器；然后从目标领域无标签样本中选择特有特征来对通用分类器进行精确化从而得到适合于目标领域数据的分类器。戴文渊（音译，W.Y. Dai）等人[4]提出了一种基于联合聚类（Co-clustering）的预测领域外文档的分类方法 CoCC。该方法通过对类别和特征进行同步聚类，实现知识与类别标签的迁移。CoCC 算法的关键思想是识别出领域内(也称为目标领域)与领域外(也称为源领域)数据共有的部分，即共有的词特征，于是，类

别信息以及知识通过这些共有的词特征从源领域传到目标领域。

2.2.4 基于特征映射方法

基于特征映射的迁移学习方法是把各个领域的数据从原始高维特征空间映射到低维特征空间，在该低维空间下，源领域数据与目标领域数据拥有相同的分布[3][63-65]。这样就可以利用低维空间表示的有标签的源领域样本数据训练分类器，对目标测试数据进行预测。该方法与特征选择的区别在于这些映射得到的特征不在原始的特征当中，是全新的特征。

潘嘉林等人[63]提出了一种新的维度降低迁移学习方法，他通过最小化源领域数据与目标领域数据在隐性语义空间上的最大均值偏差（Maximun Mean Discrepancy），求解得到降维后的特征空间。在该隐性空间上，不同的领域具有相同或者非常接近的数据分布，因此就可以直接利用监督学习算法训练模型对目标领域数据进行预测。顾全泉（音译，Quanquan Gu）等人[60]探讨了多个聚类任务的学习(这些聚类任务是相关的)，提出了一种寻找共享特征子空间的框架。在该子空间中，各个领域的数据共享聚类中心，而且他们还把该框架推广到直推式迁移分类学习。布利泽（Blitzer）等人[3]提出了一种结构对应学习算法(Structural Corresponding Learning，SCL)。该算法把领域特有的特征映射到所有领域共享的“轴”特征，然后就在这个“轴”特征下进行训练学习。结构对应学习算法已经被用到词性标注[3]以及情感分析[66]中。类似的工作还有引文[67]等。

2.2.5 基于权重方法

在迁移学习中，有标签的源领域数据的分布与无标签的目标领域数据的分布是不一样的，因此那些有标签的样本数据并不一定是全部有用的。如何侧重选择那些对目标领域分类有利的训练样本？这就是基于实例的迁移学习所要解决的问题。基于实例的迁移学习通过度量有标签的训练样本与无标签的测试样本之间的相似度来重新分配源领域中样本的采样权重。相似度大的，即对训练目标模型有利的训练样本被加大权重，否则权重被削弱。蒋静等人[23]提出了一种实例权重框架来解决自然语言处理任务下的领域适应问题。他们首先从分布的角度分析，认为产生领域适应问题的原因主要有两方面：实例的不同分布以及分类函数的不同分布。因此他们提出了一个最小化分布差异性的风险函数，来解决领域适应性问题。戴文渊等人[12] 将 Boosting 学习算法扩展到迁移学习中，提出了 TrAdaBoost 算法。在每次迭代中改变样本被采样的权重，即在迭代时降低源领域中的样本权重，加大有利于模型训练的目标领域中的样本权重。他们还用“概率近似正确”方法分析证明了该算法的有效性。下面简要介绍 TrAdaBoost 算法。

用于迁移学习任务中的源领域数据与目标领域数据虽然分布不同，但是相关的。也就是辅助的源领域中，存在一部分比较适合用来学习一个有效的分类模型的训练样本，并且这些样本对目标测试样本是适用的。于是 TrAdaBoost 算法的目标就是从辅助的源数据中找出那些适合测试数据的实例，并把这些实例迁移到目标领域中少量有标签样本的学习中去。该算法的关键思想是利用 Boosting 的技术过滤掉源领域数据中那些与目标领域中少量有标签样本相似性最差的样本数据。其中，Boosting 技术用来建立一种自动调整权重机制，于是重要的源领域样本数据权重增加，不重要的源领域样本数据权重减小。在 TrAdaBoost 中，AdaBoost[68]被用在目标领域中少量有标签样本，以保证分类模型在目标领域数据上的准确性；而 Hedge(β)[68]被用在源领域数据上，用于自动调节源领域数据的重要度。一个直观TrAdaBoost 的例子如图 2 所示。另外对参数加权组合的工作，如引文[69]。

根据是否从多个源领域数据学习，迁移学习算法又可以分为单个源领域以及多个源领域的迁移学习。本-戴维（Ben-David）等人[2]分析了领域数据的表示，并提出了一个很好的模

型。该模型不仅使分类模型在训练数据上的泛化误差最小化，而且使源领域与目标领域之间的不同性最小化。凌霄（音译，Xiao Ling）等人[70]提出了一种新的光谱分类算法，该算法通过优化一个目标函数来寻找源领域中的监督信息与目标领域的本质结构之间的最大一致性。庄福振等人[17]综合半监督学习中的几种正则化准则，提出了基于混合正则化准则的迁移学习框架。马哈茂德（Mahmud）等人[51][71]从算法信息论的角度来研究迁移学习，该方法度量了不同任务之间的相关性，然后决定多少信息可以做迁移以及怎么迁移这些信息。邢迪侃（音译，DikanXing）等人[7]提出了一种直推式迁移学习方法，该方法首先开发利用所有数据集（包括源领域数据和目标领域数据）上的几何分布结构，然后再利用目标领域上的流形结构。针对多源领域学习问题，高静（Jing Gao）等人[72]提出了一种多模型局部结构映射方案，实际上是对不同源领域训练得到的模型赋予不同的投票权重，而该权重是由预测样本本身的局部分布结构决定的。高静等人[73]解决了不同模型的一致性问题。这两个多源领域学习的工作很好地处理了多个模型的集成问题。为了更加深入地挖掘、开发各个源领域数据的内部结构或者数据分布，罗平和庄福振等人[14][18]提出了一致性正则化框架。在这个框架下，局部的子分类器不仅考虑了在源领域上的可利用的局部数据，而且考虑了这些由源领域知识得到的子分类器在目标领域上的预测的一致性。段立新（音译，Lixin Duan）等人[15]利用源领域训练得到的模型作为辅助分类器。

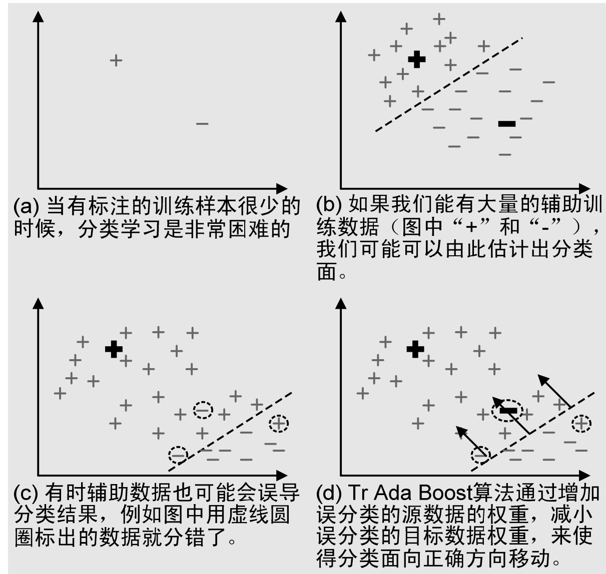


图2. 关于 TrAdaBoost 算法思想的一个直观示例[12]

3　迁移学习相关理论研究

从理论层面讲，迁移学习问题研究以下问题：第一，什么条件下从源领域数据训练出的分类器能够在目标领域表现出优异的分类性能，即什么条件下可进行迁移？第二，给定无标注目标领域，或者有少量的标记数据，如何在训练过程中与大量有标记的源数据结合使得测试时的误差最小，即迁移学习算法的研究。目前对迁移学习理论研究比较多的主要是在领域适应性方面。

关于领域适应性问题的理论分析最早是本-戴维等在文献[2]中提出的。该文基于 VC 维[2]对领域适应性问题给出了推广性的界。该文最有价值的贡献在于定义了分布之间的距离，此距离与领域适应性有关。在此基础上，对有限 VC 维情况，可用他们在引文[74]中提出的方法，从有限个样本估计适应推广能力。但是当 VC 维不是有限的情况下会有什么样的结论该文没有涉及，需要进一步探讨。另外不同的领域分布之间的距离会得出不同精度的误差估计，由此可以通过研究各具特色的距离用于解决领域适应性问题，以适应不同应用场合的需要。本-戴维[2]还通过实验指出结构对应学习方法确实能够达到 $\mathcal{A}$ 距离最小的同时间隔损失最低，从而提高目标领域上的学习性能。本-戴维等人[2]分析了领域数据的表示，并提出了一个很好的模型。该模型不仅使分类模型在训练数据上的泛化误差最小化，而且使源领域与目

2 Vapnik-Chervonenkis Dimension，是一种对统计分类算法能力的度量，详见 http://en.wikipedia.org/wiki/VC_dimension

标领域之间的不同性最小化。这项工作后续研究的阶段性成果见引文 [33]，该文从源数据加权组合获得模型，并给出在特定的经验风险最小化的情形下的误差率。最新的成果发表在 2010 年的 *Machine Learning*（《机器学习》）杂志上[75]。该文研究了在什么条件下一个分类器能在目标领域很好完成分类任务，还研究了给定目标领域少量的已标注的样本，如何在训练过程将其与大量的已标注的源数据相结合，使得目标误差最小。

曼苏尔（Mansour）指出对任意给定的目标函数，存在一个对源假设的领域加权分布组合使得损失至多为给定的值[76]。他还对于任意的目标分布，给出了基于源领域和目标领域之间的雷尼散度（Rényi divergence ）的领域推广误差[77]。更为精确的推广误差上界估计应用到回归和一般的损益函数，并提出通过加权实现经验分布更好地反映目标领域分布[78]。目前的工作尽管已经进行了一些理论尝试，但还远远不足，对迁移学习有效性的理论研究还有待进一步深入。

下面介绍下我们利用生成模型在迁移学习方面做的工作。

4 基于生成模型的迁移学习方法

目前很多迁移学习算法都是基于判别模型 [13][63-65]，判别算法是根据给定源领域数据 $\boldsymbol{X}$，直接训练得到判别模型 $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$。由于源领域与目标领域数据分布不一致，判别模型没有考虑联合概率 $\boldsymbol{P}(\boldsymbol{X},\boldsymbol{Y})$，因此有时不能得到很好的预测结果。与判别模型不同，生成模型先计算得到联合概率 $\boldsymbol{P}(\boldsymbol{X},\boldsymbol{Y})$，然后再计算 $\boldsymbol{P}(\boldsymbol{Y}|\boldsymbol{X})$。这样，就提供了一种很好的机制，可以为源领域和目标领域数据不同分布建模，实现源领域与目标领域之间的知识迁移，从而提高算法的性能[79-82]。

在迁移学习文本分类中，源领域数据与目标领域数据在原始词特征上分布不一致，也就是说它们可能会采用不同的词特征来表示同一个语义概念。但我们发现不同的领域数据，其词特征聚类(又称词特征概念)与文档类别(又称文档聚类、文档概念)之间的关联关系可能是一样的。比如，表示词特征概念“Computer Science”的词有“hardware”、“software”、“program”、“programmer”、“disks”以及“ROM”等，但是这些词在不同的领域中可能频率相差很大。在关于硬件公司的新闻网页中，“hardware”、“disks”以及“ROM”可能是高频词，而在关于软件公司的新闻网页中，“software”、“program”以及“programmer”更可能是高频词。因此不同的领域表示同一个概念的词特征差异很大，这就会导致用原始特征训练得到的分类器可能是不可靠的。如果我们能够找出各个领域的词特征概念，并用它们来预测样本的类别，那么就会比直接用原始特征要可靠和有效。从上面的例子可以看到，一个网页不管是来自于哪一个领域，只要其包含特征概念“Computer Science”，那么该网页就是属于计算机相关的文档类。我们把表示词特征概念的词，定义为词概念外延，把词特征概念与文档类别之间的关系定义为词概念内涵，文档类别中包含的具体文档定义为文档类别外延。我们[79] 研究了基于生成模型的挖掘多领域之间共性与特性的跨领域分类方法，对有效挖掘词特征聚类与文档类别关联关系进行了深入研究。其主要思想如图 3 所示。

图 3 中，每个大矩形框中又包含两个小的矩形框，分别为各个领域词特征概念的外延和文档概念的外延。领域的特性包括所有的外延，而领域的共性则是它们共享的词特征概念与文档概念之间的联合概率分布，即图中的八边形所示。实际上源领域中的数据是有标记的，即源领域中文档概念的外延已知，可以作为整个模型的监督信息，如图中的圆圈所示。这些监督信息通过领域之间的共性实现知识的迁移。领域的共性起到桥的作用，最后实现对目标领域数据的分类预测。实验结果表明该算法具有较强的迁移学习能力，可以处理迁移学习比

较难的分类问题。我们还对基于判别模型和基于生成模型的迁移学习算法进行了初步的探讨，认为基于生成模型的方法可以有效地对源领域与目标领域之间的差异进行建模，这可能更加适合迁移学习。

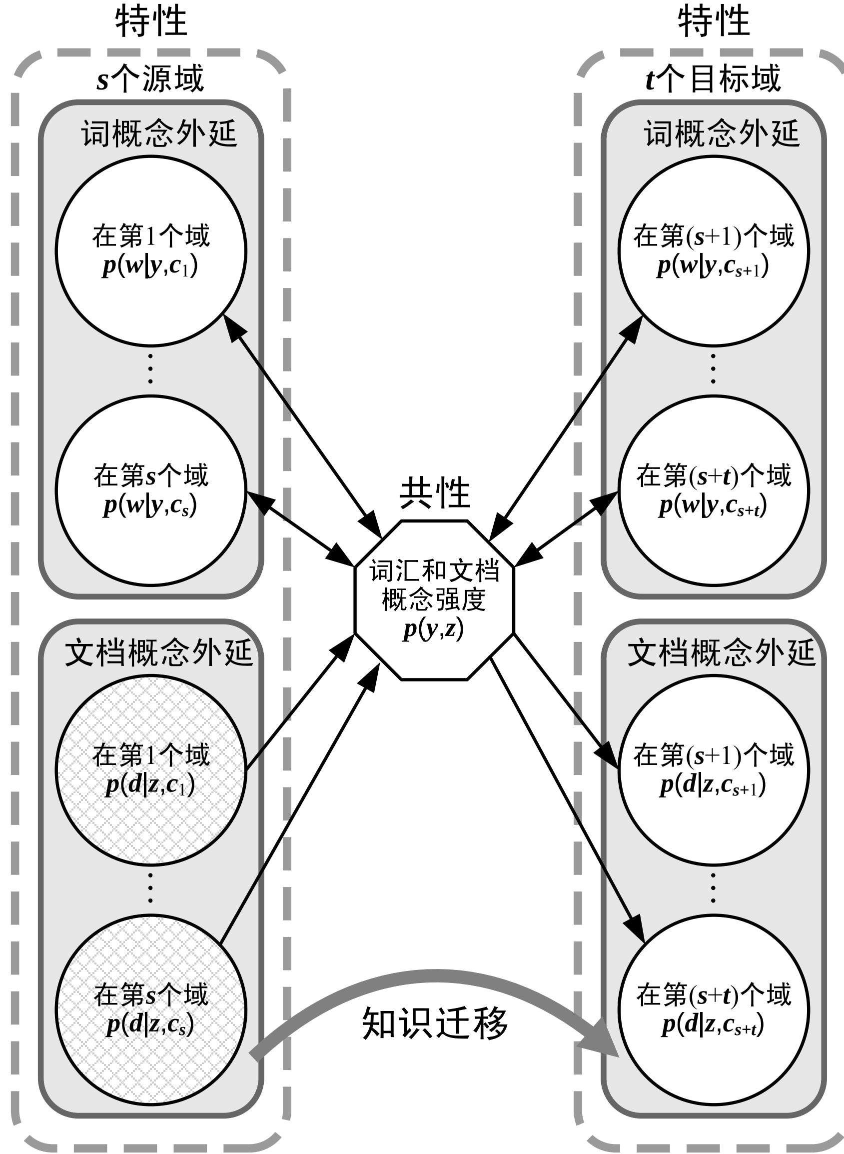


图3. 不同领域之间的共性和特性

以往的工作[79][81]假设源领域和目标领域共享相同的概念集，但是除了共享概念以外，不同领域可能还包含自己独特的概念。我们对不同领域的概念进行了深入的分析[82]，把概念分成三类：一致性概念、相似概念以及领域特有的概念。我们提出了一般的概率统计模型来挖掘这三种概念，并开发了一种期望最大化（Expectation-Maximization，EM）算法进行求解。大量的实验结果表明所提出的模型优于作为对比的迁移学习算法。

5 未来研究方向

本文系统地给出了迁移学习算法以及相关理论的研究进展。迁移学习作为一个新兴的研究领域还很年轻，目前工作主要还是集中在算法方面，因此值得我们进一步研究。

迁移学习最早来源于教育心理学，这里借用美国心理学家贾德(Judd, C.H.)[3]提出的“类化说”学习迁移理论来讨论目前机器学习领域迁移学习研究还存在的三个问题。首先，贾德认为在先期学习 A 中获得的东西，之所以能迁移到后期学习 B 中，是因为在学习 A 时获得了一般原理，这种原理可以部分或全部运用于 A、B 之中。根据这一理论，两个学习活动之间存在的共同要素是产生迁移的必要前提。这也就是说，想从源领域中学习知识并运用到目标领域中，必须保证源领域与目标领域有共同的知识。那么如何度量这两个领域的相似性与共同性，是问题之一。第二，贾德的研究表明，知识的迁移是存在的，只要一个人对他的经验、知识进行了概括，那么从一种情境到另一种情境的迁移是可能的。知识概括化的水平越高，迁移的范围和可能性越大。把该原则运用到课堂上，同样的教材采用不同的教学方法，产生的迁移效果是不一样的，既可能产生积极迁移也可能产生相反的作用。即同样的教材内容，由于教学方法不同，而使教学效果大为悬殊，迁移的效应也大不相同。所以针对不同的学习问题，研究有效的迁移学习算法也是另一个重要问题。第三，根据贾德的泛化理论，重要的是在讲授教材时要鼓励学生对核心的基本的概念进行抽象或概括。抽象与概括的学习方法是最重要的方法，即要求学生在学习的时候对知识进行思维加工，区别本质的和非本质的属性，偶然的和必然的联系，舍弃那些偶然的、非本质的东西，牢牢把握那些必然的本质的东西。这种学习方法能使学生的认识从低级的感性阶段上升到高级的理性阶段，从而实现更广泛更成功的正向迁移。也就是说在迁移学习的过程中，应该避免把非本质的、偶然的知识当成本质的(领域共享的)、必然的知识，这样才能实现正迁移。所以，如何实现正迁移，避免负迁移也是迁移学习的一个重要研究问题。

由此，我们认为后续研究有以下几个可能的方向：第一，研究领域相似性、共同性的准确度量方法；第二，除了目前很受重视的迁移学习分类算法外，其他方面的应用算法有待进一步研究，比如情感分类，强化学习，排序学习，度量学习，人工智能规划等；第三，研究迁移学习算法有效性的理论，如：可迁移学习条件，如何获取实现正迁移所需要的本质属性，如何避免负迁移；最后，在大数据环境下，研究高效的迁移学习算法尤为重要。目前的研究主要还是集中在研究领域，数据量小而且测试数据非常标准，应把研究的算法瞄准实际应用数据，以适应目前大数据挖掘研究浪潮。

参考文献：

[1] He Q, Zhao X R, Shi Z Z. (2008). Minimal Consistent Subset for Hyper Surface Classification Method. *International Journal of Pattern Recognition and Artificial Intelligence*, 22(1): 95-108.

[2] Ben-David S, Blitzer J, Crammer K, et al. (2007). Analysis of Representations for Domain Adaptation. *Proceedings. of Advances in Neural Information Processing Systems 19,* Cambridge: MIT Press: 137-144.

[3] Blitzer J, McDonald R, Pereira F. (2006). Domain Adaptation with Structural Correspondence Learning. *Proceedings. of the International Conference on Empirical Methods in Natural Language Processing*, Stroudsburg PA: ACL: 120-128.

[4] Dai W Y, Xue G R, Yang Q, et al. (2007). Co-clustering based Classification for Out-of-domain Documents. *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining,* New York: ACM Press: 210-219.

[5] Dai W Y, Xue G R, Yang Q, et al. (2007). Transferring Naive Bayes Classifiers for Text Classification. *Proceedings of 22nd Conference on Artificial Intelligence,* California 94025: AAAI Press: 540-545.

[6] Liao X J, Xue Y, Carin L. (2005). Logistic Regression with an Auxiliary Data Source. *Proceedings of 22nd International Conference on Machine Learning*, San Francisco: Morgan Kaufmann: 505-512.

[3] http://baike.baidu.com/subview/226455/11107073.htm

[7] Xing D K, Dai W Y, Xue G R, et al. (2007). Bridged Refinement for Transfer Learning. *Proceedings of 11th European Conference on Practice of Knowledge Discovery in Databases*, Berlin: Springer-Verlag: 324-335.

[8] Mahmud M M H. (2007). On Universal Transfer Learning. *Proceedings of 18th International Conference on Algorithmic Learning Theory*, Sendai, Japan: 135—149.

[9] Samarth S, Sylvian R. (2006). Cross Domain Knowledge Transfer Using Structured Representations. *Proceedings of 21nd Conference on Artificial Intelligence*, California 94025: AAAI Press: 506-511.

[10] Bel N, Koster C H A, Villegas M. (2003). Cross-Lingual Text Categorization. *Proceedings of European Conference on Digital Libraries,* Berlin: Springer-Verlag: 126-139.

[11] Zhai C X, Velivelli A, Yu B. (2004). A Cross-collection Mixture Model for Comparative Text Mining. *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM: 743—748.

[12] Dai W Y, Yang Q, Xue G R, et al. (2007). Boosting for Transfer Learning. *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann: 193-200.

[13] Pan S J, Yang Q. (2010). A Survey on Transfer Learning [J]. *IEEE Transaction on Data Engineering*, 22(10): 1345-1359.

[14] Luo P, Zhuang F Z, Xiong H, et al. (2008). Transfer Learning from Multiple Source Domains via Consensus Regularization. *Proceedings of 17th ACM Conference on Information and Knowledge Management*, New York: ACM Press: 103-112.

[15] Duan L X, Tsang Ivor W, Xu D, et al. (2009). Domain Adaptation from Multiple Sources via Auxiliary Classifiers. *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA: ACM: 289—296.

[16] Dai W Y, Chen Y Q, Xue G R, et al. (2008). Translated Learning: Transfer Learning across Different Feature Spaces. *Proceedings. of Advances in Neural Information Processing Systems 20,* Cambridge: MIT Press: 353-360.

[17] Zhuang F Z, Luo P, He Q, et al. (2009). Inductive Transfer Learning for Unlabeled Target-domain via Hybrid Regularization. *Chinese Science Bulletin*, 54(14): 2470-2478.

[18] Zhuang F Z, Luo P, Xiong H, et al. (2010). Cross-domain Learning from Multiple Sources: A Consensus Regularization Perspective. *IEEE Transactions On Knowledge And Data Engineering*, 22(12): 1664-1678.

[19] Rosenstein M T, Marx Z, Kaelbling L P. (2005). To Transfer or Not to Transfer. *Proceedings of Neural Information Processing Systems 2005 workshop on Inductive Transfer: 10 years Later*, Cambridge: MIT Press.

[20] Dai W Y, Jin O, Xue G R, et al. (2009). Eigen Transfer: a Unified Framework for Transfer Learning. *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann: 193-200.

[21] 戴文渊. (2008). 基于实例和特征的迁移学习算法研究. 上海: 上海交通大学计算机科学与工程系. 2008. 1-55.

[22] 施潇潇. (2009). 主动迁移学习模型的研究与应用. 中山: 中山大学计算机应用技术. 1-69.

[23] Jiang J, Zhai C X. (2007). Instance Weighting for Domain Adaptation in NLP. *Proc. of 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg PA: *Association for Computational Linguistics*: 264-271.

[24] Lee S, Chatalbashev V, Vickrey D, et al. (2007). Learning A Meta-level Prior for Feature Relevance from Multiple Related Tasks. *Proceedings of the 24th International Conference on Machine Learning*, New York, NY, USA: ACM: 489—496.

[25] Wang C, Mahadevan S. (2008). Manifold Alignment Using Procrustes Analysis. *Proceedings of 25th*

International Conference on Machine Learning, New York, NY, USA: ACM: 1120—1127.

[26] Lawrence N D, Platt J C. (2004). Learning to Learn with the Informative Vector Machine. *Proceedings of the 21st International Conference on Machine Learning*, New York, NY, USA: ACM: 65—72.

[27] Schwaighofer A, Tresp V, Yu K. (2005). Learning Gaussian process kernels via hierarchical Bayes. *Proceedings of Advances in Neural Information Processing Systems 17*, Cambridge: MIT Press: 1209—1216.

[28] Evgeniou T, Pontil M. (2004). Regularized Multi--task Learning. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA: ACM: 109—117.

[29] Zadrozny B. (2004). Learning and Evaluating Classifiers under Sample Selection Bias. *Proceedings of the Twenty-first International Conference on Machine learning*, New York, NY, USA: ACM: 114—121.

[30] Huang J Y, Smola A J, Gretton A, et al. (2007). Correcting Sample Selection Bias by Unlabeled Data. *Proceedings. of Advances in Neural Information Processing Systems 19,* Cambridge: MIT Press: 601-608.

[31] Fan W, Davidson I, Zadrozny B, et al. (2005). An Improved Categorization of Classifier's Sensitivity on Sample Selection Bias. *Proceedings of the 5th International Conference on Data Mining*, Los Vaqueros: IEEE Computer Society: 605—608.

[32] Ando R K, Zhang T. (2005). A high-performance semi-supervised learning method for text chunking. *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA: 1—9.

[33] Blitzer J, Crammer K, Kulesza A, et al. (2008). Learning Bounds for Domain Adaptation. *Proceedings. of Advances in Neural Information Processing Systems 20,* Cambridge: MIT Press : 129-136.

[34] Dai W Y, Yang Q, Xue G R, et al. (2008). Self-taught Clustering. *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann: 200-207.

[35] Wang Z, Song Y Q, Zhang C S. (2008). Transferred Dimensionality Reduction. *Proceedings of the European conference on Machine Learning and Knowledge Discovery in Databases*, Berlin, Heidelberg: Springer-Verlag: 550—565.

[36] Zhou Z H. (2006). Learning with Unlabeled Data and Its Application to Image Retrival. *In Proc. of 9th Pacific Rim International Conference On Artificial Intelligence*, Berlin: Springer-Verlag: 5-10.

[37] Zhu X J. (2005). Semi-supervised Learning Literature Survey. Department of Computer Sciences, University of Wisconsin, Madison.

[38] Joachims T. (1999). Transductive Inference for Text Classification Using Support Vector Machines. *Proceedings of 16th International Conference on Multimedia,* Augsburg Germany, New York: ACM Press: 200-209.

[39] Joachims T. (2003). Transductive Learning via Spectral Graph Partitioning. *Proceedings of 16th International Conference on Multimedia,* Augsburg Germany, New York: ACM Press: 290-297.

[40] Tong S, Chang E. (2001). Support Vector Machine Active Learning for Image Retrieval. *Proceedingss of 9th ACM International Conference on Multimedia*, New York: ACM Press: 107-118.

[41] Cohn D, Atlas L, Ladner R. (1994). Improving Generalization with Active Learning. *Machine Learning,* Berlin: Springer-Verlag, 15(2): 201-221.

[42] Sindhwani V, Niyogi P. (2005). A Co-regularized Approach to Semi-supervised Learning with Multiple Views. *Proceedings of the ICML Workshop on Learning with Multiple Views*, San Francisco: Morgan Kaufmann: 74-79.

[43] Yarowsky D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, Stroudsburg, PA, USA: Association for Computational Linguistics: 189—196.

[44] Blum, Mitchell T. (1998). Combining Labeled and Unlabeled Data with Co-training. *Proceedings of the eleventh annual conference on Computational learning theory*, New York, NY, USA: 92—100.

[45] Dasgupta S, Littman M L, Mcalleste D. (2001). PAC Generalization Bounds for Co-Training. *Proceedings. of Advances in Neural Information Processing Systems 13,* Cambridge: MIT Press: 375-382.

[46] Abney A. (2002). Bootstrapping. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, USA: 360-367.

[47] Abney A. (2004). Understanding the Yarowsky Algorithm. *Journal of Computational Linguistics*, Cambridge, MA, USA: MIT Press: 365—395.

[48] Shi X X, Fan W, Ren J T. (2008). Actively Transfer Domain Knowledge. *Proceedings of The European Conference on Machine learning and Knowledge Discovery in DataBases,* Berlin: Springer-Verlag: 342-357.

[49] Belkin M, Niyogi P, Sindhwani V. (2007). Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research*, (7): 2399-2434.

[50] Grandvalet Y, Bengio Y. (2005). Semi-supervised Learning by Entropy Minimization. *Proceedings. of Advances in Neural Information Processing Systems 17,* Cambridge: MIT Press: 529-536.

[51] Mann G S, McCallum A. (2007). Simple, Robust, Scalable Semi-supervised Learning via Expectation Regularization. *Proceedings of 24th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann: 593-600.

[52] Raina R, Battle A, Lee H, et al. (2007). Self-taught Learning: Transfer Learning from Unlabeled Data. *Proceedings of 24th International Conference on Machine Learning,* San Francisco: Morgan Kaufmann: 759-766.

[53] Caruana R. (1997). Multitask Learning. *Machine Learning of Special issue on Inductive Transfer*, Berlin: Springer-Verlag, 28(1): 41-75.

[54] S. Thrun. (1996). Is learning the *n*-th Thing Any Easier Than Learning The First? *Proceedings of Advances in Neural Information Processing Systems 8*, Cambridge: MIT Press: 640-646.

[55] Bakker B, Heskes T. (2003). Task clustering and gating for bayesian multitask learning. *The Journal of Machine Learning Research*, 4:83-99.

[56] Bai J, Zhou K, Xue G R, et al. (2009). Multi-task Learning for Learning to Rand in Web Search. *Proceedings of 18th ACM Conference on Information and Knowledge Management*, New York: ACM Press: 1549-1552.

[57] Jebara T. (2004). Multi-task Feature and Kernel Selection for SVMs. *Proceedings of 21th International Conference on Machine Learning*, San Francisco: Morgan Kaufmann: 55-62.

[58] Argyriou A, Evgeniou T, Pontil M. (2007). Multi-task Feature Learning. *Proceedings of Advances in Neural Information Processing Systems 19*, Cambridge: MIT Press: 243-272.

[59] Obozinski G, Taskar B, Jordan M I. (2006). Multi-task Feature Selection. Department of Statistics, University of California, Berkeley.

[60] Gu Q Q, Zhou J. (2009). Learning the Shared Subspace for Multi-Task Clustering and Transductive Transfer Classification. *Proceedings of the 9th International Conference on Data Mining*, Los Vaqueros: IEEE Computer Society: 159-168.

[61] Jiang J, Zhai C X. (2007). A two-stage approach to domain adaptation for statistical classifiers. *Proc. of 16th ACM Conference on Information and Knowledge Management*, New York: ACM Press:

401-410.

[62] Zhuang F Z, Luo P, Xiong H, et al. (2010). Exploiting Associations between Word Clusters and Document Classes for Cross-domain Text Categorization. *Proceedings of the 10th SIAM Conference on Data Ming*, Philadelphia: SIAM Press: 13-24.

[63] Pan S J, Kwok J T, Yang Q. (2008). Transfer Learning via Dimensionality Reduction. *Proceedings of the 23rd Conference on Artificial Intelligence*, California 94025: AAAI Press: 677-682.

[64] Si S, Tao D C, Chan K P. (2010). Evolutionary Cross-domain Discriminative Hessian Eigenmaps [J]. *IEEE Transactions on Image Processing*, 19(4): 1075-1086.

[65] Si S, Tao D C, Geng B. (2010). Bregman Divergence-based Regularization for Transfer Subspace Learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7): 919-942.

[66] Blitzer J, Dredze M, Pereira F. (2007). Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classification. *Proc. of 45th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg PA: ACL: 440-447.

[67] Xie S H, Fan W, Peng J, et al. (2009). Latent Space Domain Transfer between High Dimensional Overlapping Distributions. *Proceedings of ACM Conference on World Wide Web*, New York: ACM Press: 91–100.

[68] Freund Y, Schapire R E. (1997). A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *Journal of Computer and System Sciences*, 55(1): 119-139.

[69] Dredze M, Kulesza A, Crammer K. (2010). Multi-domain learning by confidence-weighted parameter combination. *Journal of Machine Learning*, 79(1-2): 123-149.

[70] Ling X, Dai W Y, Xue G R, et al. (2008). Spectral Domain-Transfer Learning. *Proceedings of 14th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press: 488-496.

[71] Mahmud M M H, Ray S. (2007). Transfer Learning Using Kolmogorov Complexity: Basic Theory and Empirical Evaluations. *Proc. of Advances in neural information processing systems*, MIT Press: 985-992.

[72] Gao J, Fan W, Jiang J, et al. (2008). Knowledge Transfer via Multiple Model Local Structure Mapping. *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press: 283-291.

[73] Gao J, Fan W, Sun Y Z, et al. (2009). Heterogeneous Source Consensus Learning via Decision Propagation and Negotiation. *Proceedings of 13th ACM International Conference on Knowledge Discovery and Data Mining*, New York: ACM Press: 339-348.

[74] Kifer D, Ben-David S, Gehrke J. (2004). Detecting Change in Data Streams. *Proceedings of the 30th International Conference on Very Large Data Bbases*, Toronto, Canada: VLDB Endowment: 180—191.

[75] Ben-David S, Blitzer J, Crammer K, et al. (2010). A theory of learning from different domains. Journal of Machine Learning, 79(1-2): 151—175.

[76] Mansour Y, Mohri M, Rostamizadeh A. (2008). Domain Adaptation with Multiple Sources. *Proceedings of Advances in Neural Information Processing Systems 20*, Cambridge: MIT Press: 1—8.

[77] Mansour Y, Mohri M, Rostamizadeh A. (2009). Multiple source adaptation and the R\ényi divergence. *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, Arlington, Virginia, United States: AUAI Press: 367—374.

（下转第 12 页）